Recovering approximate single cell distribution from aggregate measurements

Pratik Worah PWORAH@GOOGLE.COM
Google Research, USA

Abstract

We design an algorithm that approximately recovers single cell data, for example, scRNA-seq data for a small subset of genes, using aggregate measurements, for example, unlabeled in-situ hybridization data and RNA-seq data, from a sample. The technical crux of our algorithm involves compressed sensing based sparse recovery when the measurement matrix is unknown, only its statistical distribution is known.

1. Introduction

Since the early 2000s, compressed sensing based sparse recovery has grown to prove its usefulness in several domains, ranging from communication, to medicine, to image processing, and even aviation (see Chapter 1 in the book Foucart and Rauhut (2013) for a long list of references). The basic paradigm is simple: if we know the underlying signal is sparse then we can recover it using only a few measurements, as long as the measurements are chosen carefully. In this paper, we detail one such application to approximately recover single cell RNA distributions from few measurements.

The importance of single cell data is increasing as we try to understand fundamental processes in biology and apply them in medicine. However, the costs of gathering single cell data are prohibitive, so a method to even approximately recover single cell data from a small number of inexpensive aggregate level measurements may prove useful. In this paper, we explore one such idea using compressed sensing type algorithms. We show, using simulations, that it is possible to approximately recover the underlying single cell distribution of RNA expression using ℓ_1 minimization, even when the problem does not fit the standard compressed sensing paradigm.

To summarize our results: From a healthcare and medicine perspective, our contribution is encapsulated in Algorithm 1 which uses compressed sensing without knowing the exact measurement matrix, just its statistical properties, to reconstruct the original microscopic distribution using only a few aggregate measurements of the sample. For lack of data and resources, we don't run experiments, but we run simulations (in Section 4) using publicly available scRNA-seq data (Bastidas-Ponce et al. (2019); Bergen et al. (2020))[1] as the ground truth, to see how Algorithm 1 and its derivatives may eventually perform in the real world. From a theoretical perspective, our contribution is encapsulated in Theorems 7 and 8. The

[1] Their data is available at https://scvelo.readthedocs.io/en/stable/scvelo.datasets.pancreas/ and https://github.com/theislab/pancreatic-endocrinogenesis/

latter is a generalization of a state of the art result in compressed sensing for Gaussian matrices by Raskutti et al. (2010).

Generalizable Insights about Machine Learning in the Context of Healthcare

From a machine learning perspective, Algorithm 1 extends the basic compressed sensing algorithm, which assumes that the measurement matrix is known, to a setting where the measurement matrix is unknown and only its statistical properties are known. This leads to worse recovery guarantees, but nevertheless, we prove an upper-bound on the reconstruction error in Theorem 7. In order to improve upon the reconstruction error, we extend the work of Raskutti et al. (2010) in Theorem 8 to optimize the design of the underlying measurement procedures as well. In the context of healthcare and medicine, compressed sensing algorithms have found use from radiology (see for example Jaspan et al. (2015)) to transcriptomics (see for example Cleary et al. (2021); Chang et al. (2014)). In each of those cases, the measurement matrix is assumed known, but it is entirely possible that the measurement matrix is too noisy or even unknown in some variation of those problems. The averaging approach in Algorithm 1, together with a more precise model of their measurement methods, may be helpful in solving any such cases, as they arise.

2. Related Work

2.1. Compressed sensing: Basics

A basic problem in the field of compressed sensing is to reconstruct an unknown sparse[2] signal $x \in \mathbb{R}^n$ using data from a (small) sample of measurements. These measurements are represented by a matrix $\Phi \in \mathbb{R}^{m \times n}$, where m is the number of samples. Typically, $m \simeq \log^{O(1)} n$. Moreover, the m sample observations y are related to x as: $\Phi x = y$ or in the case of noisy measurements: $\|\Phi x - y\|_2 \leq \eta$. The matrix Φ is assumed to be known. Moreover, each co-ordinate of Φx represents a weighted average of the co-ordinates of the unknown x. Therefore, these measurements are bulk or average measurements – for example, the average insulin concentration in a unit volume of blood, or the concentration of hybridized RNA in a tissue sample after the individual cellular compartments have been removed.

The central problem in compressed sensing is to reconstruct (or recover) the s-sparse unknown signal x (for $s \ll n$) from y, while keeping $m \ll n$ (see Foucart and Rauhut (2013)). The crux of the solution lies in solving a convex program (CP) of the form:

$$\begin{aligned} \min_x \quad & \|x\|_1 \\ s.t. \quad & \|\Phi x - y\|_2 \leq \eta. \end{aligned} \tag{1}$$

Definition 1 *A matrix Φ has the (s, γ)-null space property (NSP) if*

$$\forall v \in \mathrm{Ker}(\Phi) \setminus \{0\}, \ \forall |T| \leq s, \|v_T\|_1 \leq \gamma \|v_{T^c}\|_1.$$

While the number of fundamental results in compressed sensing are too numerous to list here, we point the reader who is new to the area to the book Foucart and Rauhut (2013),

[2] A n-dimensional real vector is said to be s-sparse if it is non-zero on at most s co-ordinates.

and the short paper Candès (2008) to get a flavor of the theoretical results in this subarea of mathematics and optimization. Below we state a canonical theorem from compressed sensing literature that we will reuse later in the paper.

Theorem 2 *Chen (2012); Cahill et al. (2016) Let $\mathcal{S}$ be the set of s-sparse vectors in $\mathbb{R}^n$. If Φ has (s, γ)-NSP, then any minimizer $\hat{x}$ of CP in Equation 1 satisfies*

$$\|\hat{x} - x\|_1 \leq \frac{4\sqrt{2}\sqrt{n}}{(1-\gamma)\lambda(\Phi)}\eta + \frac{4(1+\gamma)}{\sqrt{2}(1-\gamma)}\sigma_s(x), \tag{2}$$

where x is the original signal to be recovered, λ is the smallest positive singular value of Φ and $\sigma_s(x) := \min_{z\in\mathcal{S}} \|x - z\|_1$.

Definition 3 *Raskutti et al. (2010) Suppose the rows of a $m \times n$ matrix Φ are chosen according to independent multivariate normal distributions with $n \times n$ covariance matrix Υ. Then Φ is said to follow a (s, α, γ)-restricted eigenvalue (RE) condition of order γ, if $\|\Upsilon^{1/2}v\|_2 \geq \alpha\|v\|_2$, for all v that satisfy a (s, γ)-NSP with respect to Φ.*

Note that a RE condition of order (s, α, γ), for any constant α, implies a (s, γ)-NSP. Hence the former is stronger than the latter.

2.2. Measurements and assays: Basics

For the purposes of this article, a typical genomic assay can fall somewhere on the granularity spectrum of measuring single cell genomic data, for example, scRNA-seq or fluorescence in-situ hybridization (FISH) on the one hand; to mapping just the aggregate average value, for example, RNA-seq, on the other hand. Clearly, the former capture more information than the latter, as they measure the entire distribution of gene expression in a tissue, as opposed to just the average or total gene expression. FISH, in particular, has already found some use in healthcare (see for example Hu et al. (2014)). At the same time methods like FISH have practical drawbacks. For example, FISH requires the presence of labels (fluorescence markers) which can themselves interfere with cellular processes, it also requires elaborate sample preparation and fixation in order to measure the fluorescence intensity (see Haroon et al. (2013) for the details involved in a one-off in-solution FISH).

The assay implied by Algorithm 1, and its optimized derivative in Section 3.3, does rely on in-situ hybridization, but it does not require fluorescent markers or labels. Moreover, since it does not rely on measuring intensity, it should not require elaborate sample preparation like FISH. All it requires is that the hybridization take place first in the cellular compartments and then cellular compartments be made permeable (say via detergents) to perform RNA-seq to generate an observation – a co-ordinate for the input vector y to Algorithm 1.

The high-level idea is that a small amount of the statistical information of the single cell distribution is transferred to the aggregate measurement via the hybridization reaction in individual cellular compartments, see Assumption 5 for formal details. Therefore, if we make enough independent measurements then we may be able to approximately recover the entire distribution from aggregate measurements only, in certain circumstances.

3. Technical overview of our results and methods

3.1. Relationship between assays and compressed sensing

Suppose we have a set of noisy measurements $C := \{c_1, ..., c_N\}$ corresponding to N ”cells”. For example, the measurements could be the (spliced and unspliced) RNA concentrations corresponding to Cpe (Carboxypeptidase E[3]) in a sample of $N = 250$ cells from the pancreatic scRNA-seq data by Bastidas-Ponce et al. (2019). Suppose further that we bin (i.e., partition and round) the concentrations into n intervals. For example, for the Cpe sample, set $n = 100$, which gives: for values varying from $M_1 := 0$ to $M_2 := 35$ – the minimum and maximum values in our sample, 100 bins (intervals) of width 0.35 each. Let $x_{M_1}, x_{M_1+1*0.35}, ..., x_{M_2}$ denote the number of cells with measurement values $M_1, M_1 + 1 * 0.35, ..., M_2$ respectively. Thus the vector x represents the (non-normalized) distribution of the measurement values. Note that tuple of measurements C and the distribution x are unknown, and the goal of this paper is to approximate x using few measurements (i.e., small m).

While measurements, in this case RNA concentrations, can take on a continuum of values, they typically tend to cluster into a few values, in any snapshot, for informative genes. Figure 1 illustrates this fact. Out of 4000 genes (post-filtering) in the murine pancreatic data-set of Bastidas-Ponce et al. (2019), fewer than 5% of them had more than 10% of co-ordinates values set as non-zero when measured up to two significant figures of accuracy. Therefore, we may think of x as a sparsely supported distribution, likely characterizing various cell types.

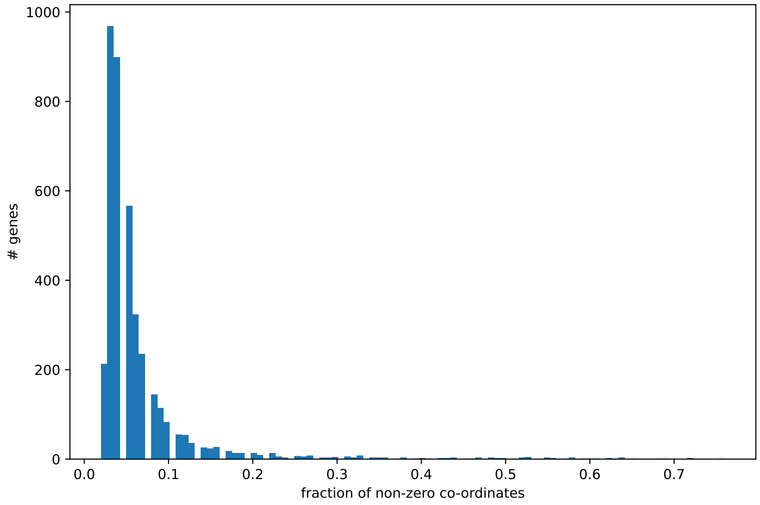


Figure 1: RNA expression is clustered into few values for most genes

Assumption 4 *We assume that the unknown distribution to be recovered, i.e., x, is s-sparse, for some $s \simeq \log^{O(1)}(n)$.*

[3] Cpe is a gene involved in the synthesis of neuropeptides and peptide hormones.

The critical question to consider is the nature of the noise in the measurement set C. It's critical because the noise distribution captures a small amount of statistical information in each cellular compartment that will be passed on to the aggregate measurement. Here's where the biochemistry of hybridization plays a role: For a cell i, the measurement c_i above (say the concentration for Cpe) would be made by an in-situ hybridization oligonucleotide probe. Note that we don't need to observe the value of c_i (for any i). All that is needed is that the excess amounts of probe are present in the cell i, so that fraction of the probe proportional to c_i hybridizes with Cpe RNA, in cell i. Of course, not all probes will hybridize. Moreover, most probes will hybridize independently of one another, and therefore one can expect roughly Gaussian statistics. Therefore, we will work under the following assumption:

Assumption 5 *We assume that the mean fraction of probes in any cell i that hybridize will be proportional to its cellular concentration c_i, the proportionality constant depending on the forward and backward hybridization reaction rates (of Cpe in our running example). Moreover, the variance will also be proportional to the concentration (of Cpe in our running example). Furthermore, we assume that the distribution of $c_1, ..., c_N$ is an independent multinomial (equivalently Gaussian, for our purposes) with mean vector $\{k \cdot c_1, ..., k \cdot c_N\}$ and variance $\{k' \cdot c_1, ..., k' \cdot c_N\}$, for proportionality constants k and k' that can be estimated empirically.*

Finally, we repeat the noisy measurements above m times, i.e., we take m samples $C(1), ..., C(m)$ and we assume that we can observe the sums: $\{\sum_{i \in [N]} c_i(j) : j \in [m]\}$ for $C(1), .., C(m)$.[4] The result is a m-dimensional vector of observations y. The sums denote a simple aggregate measurement, which tell us what total fraction of RNA probes were hybridized. Note that the information about the distribution x is now present in the mean and the variance of each of the m observations in y.

In more mathematical terms, we have the following compressed sensing problem at hand:

$$(V + \Gamma)x = y, \tag{3}$$

where V is a $m \times n$ matrix corresponding to the means i.e.,

$$V = \begin{bmatrix} M_1 & M_1 + 1 & M_1 + 2 & \dots \\ \vdots & \vdots & \ddots & \\ M_1 & M_1 + 1 & & M_2 \end{bmatrix}$$

where we have assumed that M_1 and M_2 are the minimum and maximum values of the c_i; and Γ is a matrix of mean zero independent binomial random variables corresponding to the noise and captures the variances. Thus, by Assumption 5, Γ is chosen (by nature) from a Gaussian random ensemble of the form:

$$\begin{bmatrix} N(0, k'M_1) & N(0, k'(M_1 + 1)) & N(0, k'(M_1 + 2)) & \dots \\ \vdots & \vdots & \ddots & \\ N(0, k'M_1) & N(0, k'(M_1 + 1)) & & N(0, k'M_2) \end{bmatrix}$$

[4] Notation: $[N] \equiv \{1, 2, .., N\}$.

where $N(0, \sigma^2)$ denotes a Normal random variable with mean 0 and variance σ^2, and for simplicity we have assumed a partition of $[M_1, M_2]$ into bins of width 1. Note vector y is known (observations) and x (distribution) is unknown. Moreover, while V is known, the values of the noise matrix Γ are unknown – this differentiates our problem from a standard compressed sensing problem.

3.2. Compressed sensing with unknown measurement matrix

Suppose we replaced Γ by an independent sample with the same statistics. So let $\tilde{\Gamma}$ be a sampled noise matrix and Γ the actual noise implied by the experiment. Both have the same distribution by construction, but they are independent of each other. Similarly, let x be the actual probability distribution and $\tilde{x}$ be a probability distribution that is the solution of Equation 4. Suppose $\tilde{x} = x + \Delta$ for some Δ. Then we have from our problem set-up:

$$(V + \tilde{\Gamma})\tilde{x} = (V + \Gamma)x \tag{4}$$

$$(V + \tilde{\Gamma})(x + \Delta) = (V + \Gamma)x \tag{5}$$

$$(V + \tilde{\Gamma})\Delta = \Gamma x - \tilde{\Gamma} x. \tag{6}$$

Since x is s-sparse and each row of Γ and $\tilde{\Gamma}$ consists of independent Gaussians, the Hoeffding bound (see for example Dembo and Zeitouni (1998)) implies the following observation.

Lemma 6 *The ℓ_2 norm of $(V + \tilde{\Gamma})\Delta$, i.e., the magnitude of η, is $O(ms)$, where the constants in the $O(\cdot)$ notation depends on the range of our measurements.*

3.2.1. Algorithm

Lemma 6 implies that we may solve the following compressed sensing problem to recover x:

$$\tilde{\Gamma} x = y = \Gamma x + \eta, \tag{7}$$

where the ℓ_2 norm of the noise term ϵ is $O(ms)$, and we require that $x \geq 0$. Here we have used a (known) matrix $\tilde{\Gamma}$ with a similar distribution as the (unknown) matrix Γ. Note that the bound by Raskutti et al. (2010) shows that $m = O(\log n)$ samples suffice to reconstruct the sparse vector in the problem formulated thus far. If we have several instances of the problem where $\tilde{\Gamma}$ and Γ are drawn repeatedly and independently then it is natural to expect that averaging over the solution of several such instances may lead to a reasonable approximation of the (unknown) x.

This discussion prompts the averaging method in Algorithm 1. In the algorithm, we assume that M_1 and M_2 denote the minimum and maximum values of the observations are known.

3.2.2. Upper bound on the error

Theorem 7 provides a bound on the reconstruction error of Algorithm 1.

Theorem 7 *If x is s-sparse, the error between the returned probability distribution $\tilde{p}$ and the actual distribution, denoted by p, can be bounded as:*

$$W_1(\tilde{p}, p) \leq \|\tilde{p} - p\|_1 \leq \left(\frac{4\sqrt{2}\sqrt{N}}{(1-\gamma)\lambda(\Phi)} \right) \eta, \tag{8}$$

Algorithm 1 Approximate reconstruction of distribution

```
1: Input: Measurement vector $y \in \mathbb{R}^m$, consisting of $m$ (noisy) aggregate observations, a guess of the sparsity $s$ of the unknown distribution to be found, $k$ – the averaging parameter, and $\eta$ – a parameter that upper bounds the $\ell_2$ norm of the noise.
2: Output: An approximation of the underlying $n$-dimensional probability distribution that generated $y$.
3: ▷ Begin algorithm:
4: Center the observations $y$ around 0 (by subtracting the mean $\bar{y}$ from each co-ordinate). Set $y' = y - (\bar{y}, .., \bar{y})$.
5: Sample a $m \times n$ random Gaussian ensemble $\Gamma$. The variance matrix of each row being a diagonal matrix, with the diagonal: $(vM_1, v(M_1 + 1), ..., vM_2)$. Here $v$ is a constant that is set so that the total variance of each row equals the variance of the co-ordinates of $y$.
6: The constraint in the compressed sensing problem is: $\|\tilde{\Gamma} z - y'\|_2 \leq \eta$. Solve the convex program to recover $z$.
7: Sort all coordinates of $z$ in inverse order, and set all but first $s$ to 0, to obtain $z_s$.
8: Repeat steps 3, 4 and 5 $k \simeq \log n$ times; compute $\frac{z_s(1)+...+z_s(k)}{k}$, zero out all but its $s$ largest values, normalize its mass to 1 so the resulting $\tilde{p}$ is a probability distribution supported on $s$ points and return it.
```

where $\eta = O(ms)$, *with high probability. Here* W_1 *denotes the* 1*-Wasserstein distance between* p *and* $\tilde{p}$*.*

The proof of Theorem 7 simply averages Equation 2 after plugging in the expression for the noise in Equation 6. The proof is given in the appendix.

Remark: While there is not much provable gain in the RHS of Equation 8 compared to Equation 2, the LHS of the equation can in principle be much smaller, especially when the variance of $\|x - z_s\|_1$ is large, simply by the definition of 1-Wasserstein distance. Moreover, the Wasserstein metric makes more sense for our use case, since the ℓ_1 metric does not capture the similarity in two gene expression values, if they are close-by but not exactly equal. Hence our choice of the Wasserstein metric for empirical evaluation.

Remark: Note that the upper-bound on the error increases with m, so from the perspective of bounding reconstruction error, we should use as few samples as possible while maintaining the RE condition with high confidence.

3.3. Optimizing experiment design

When we actually implement and run Algorithm 1 on a sample of the pancreatic scRNA-seq data by Bastidas-Ponce et al. (2019), we find the reconstruction error to be large. For the gene Cpe with $N = 250$, $n = 100$, $m = 10$, $s = 20$, we find the recovered distribution deviates widely from the original distribution. See Figure 2.

The reason seems to be the choice of measurement function – we use only linear measurements that count the the total number of hybridized RNA segments in each cell compartment and sums them up over all compartments. Although, the noise makes the matrix rows unequal, the minimum positive singular value of such a matrix is still quite small,

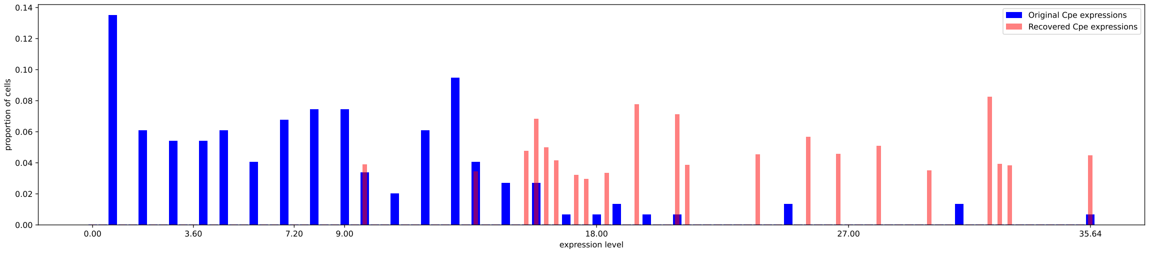


Figure 2: Recovered and original distributions for non-optimized experiment design.

which makes the upper-bound on the error, in Equation 2, weak. The way around it is to use diverse non-linear functions in our measurements. This requires solving two new problems:

1. On the experiment side: We need modifications to the measurements that are non-linear. For example, introduce probes so that a single probe can bind to two Cpe RNA segments simultaneously in the cell compartments, and then eventually count the total number of such ”doubly” hybridized probes. This generates a measurement corresponding to $\sum_{i\in[N]} c_i^2$ as opposed to $\sum_{i\in[N]} c_i$ – a quadratic (non-linear) measurement. Since we assume a Gaussian noise model where the noise (variance) is proportional to the concentration of the product (see Assumption 5), the corresponding variances will also scale as $\sum_{i\in[N]} c_i^2$. There's nothing unique about quadratics, other different non-linear function based probes can be designed and used.

2. On the compressed sensing theory side: We need to characterize the optimal choice of non-linear functions that can reduce the number of measurements m, and therefore the reconstruction error[5]. In particular, it requires a version of the upper-bound in Raskutti et al. (2010) extended to Gaussian ensembles with unequal distribution among rows.[6]

Following (1) above, suppose that $\mathcal{F}$ denotes the set of non-linear functions that can be successfully implemented using hybridization probes – *permissible non-linear functions*. Then following (2) above, Algorithm 1 can be modified so that the rows of measurement matrices Γ are picked according to a set of Normal distributions with variances specified by functions in $\mathcal{F}$. Let $\mathcal{M}$ be the set of such *permissible covariance matrices*. See also the discussion in Section 4.

While we do not have the resources to design and experiment with our own custom in-situ hybridization probes, we are able to run simulations on scRNA-seq data from Bastidas-Ponce et al. (2019), based on reasonable assumptions. We do note a marked improvement over Figure 2. See Figure 3, for example, which used simple non-linear functions of low

[5] See Theorem 7 for the relationship between m and reconstruction error.

[6] It is worth noting here, that the general problem of upper-bounding the number of samples required for reconstruction with high probability, for measurement matrices consisting of arbitrary covariance structure, has not been satisfactorily solved (see introduction of Foucart and Rauhut (2013) for the difficulties involved).

degree that should be reproducible by appropriately designed hybridization probes. Details and further simulations are presented in Section 4.

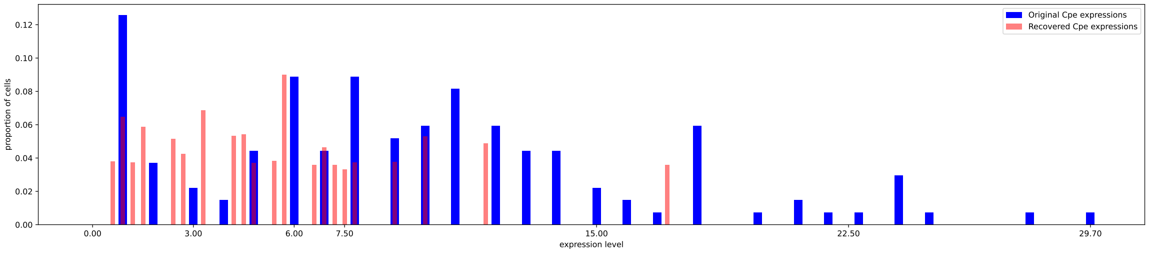


Figure 3: Recovered and original distributions using ℓ_1 minimization algorithm with a better optimized experiment design for Cpe ($N = 250, n = 100, m = 10, s = 20$).

Remark: A similar upper bound on reconstruction error as Theorem 7 holds for this case where we use different permissible covariance matrices for choosing our measurement matrix. The exact constants in the RE condition will depend on the non-linear functions used (see proof of Theorem 7).

We also show a theoretical result (Theorem 8) – an upper bound on the number of samples required for RE condition to hold in the case of the optimized measurement matrices as above. Recall from Theorem 7 that the upper-bound on the reconstruction error is $O(ms)$, so minimizing m while maintaining the NSP or RE condition will allow us to use Equation 2 to upper-bound the reconstruction error – thus the relevance of our theorem.

Theorem 8 *The optimal choice of non-linear functions (equivalently in-situ hybridization measurements) is such that it minimizes the maximum over all columns of the sum of all variances in the column.*

More formally, let $\Upsilon_1, ..., \Upsilon_m$ be the covariance matrices, chosen from a permissible set of covariance matrices $\mathcal{M}$. Row i in the $m \times n$ measurement matrix is a mean zero Gaussian with covariance matrix Υ_i. Then the optimal choice of Υs that minimizes m while the (s, α, γ)-RE condition is satisfied for some s and constants α and γ, with high probability, optimizes the function:

$$\min_{\Upsilon_1,...,\Upsilon_m} \max_j \sum_{i \in [m]} \Upsilon_i(i, j). \tag{9}$$

Our proof of Theorem 8 is a generalization of that in Raskutti et al. (2010) and requires a modification of the proof for a part of their Theorem 1 (their main result). Their result and our proof are stated in the Appendix (Subsection 6.1).

Remark: In practice, we don't need to exactly optimize the choice of the parameters, for the theoretical insights to be useful. The results in Figure 3 were produced by a simple intuition that spreading out the variances across columns should likely reduce the reconstruction error. This is because combining Theorems 7 and 8 shows that having some columns with much higher variance than others leads to a much larger upper-bound on the error in recovered vs actual distribution.

4. Simulations on pancreatic scRNA-seq data

In this section, we provide details of our simulation and a few more examples of recovered distributions with some error estimates as well.

All the plots in this paper use a uniform random sample of $N = 250$ cells from the $\simeq 4000$ cells in the original data Bastidas-Ponce et al. (2019). We assume $n = 100$, i.e., range of expression values for any single gene (say Cpe) are rounded into 100 bins, for computational tractability reasons we keep n small. The original distribution of a gene like Cpe for our sample of 250 cells would then be the histogram of: number of cells/N (y-axis) vs (rounded) gene expression value (x-axis). Recall that, from Figure 1, most genes will have sparse histograms i.e., they typically take only a few different values in our snapshot. Therefore, we assume sparsity parameter $s = 20$ in Algorithm 1, for our simulations. Our goal is to reconstruct the actual histogram using Algorithm 1 or its derivative (with optimal or near optimal variance matrix), using as few samples, i.e., m, as possible. Typically, we use a value of $m \simeq 20, k = 20$ and $\eta = 0.5$[7] in the plots below.

Genes like Cpe, Cck, Nnat play a role in secretion so they are 0 valued for non-secretory cells, and the latter as a group are far more numerous than any individual subgroup of secretory cells. So they would completely overwhelm all other bars/values in the histogram. Therefore, we remove the bucket corresponding to 0 value from our input. In any actual practice, we would need to separate out secretory and non-secretory cells before we ran an algorithm like ours.

Finally, we reconstruct the distribution for one gene at any time, for computational time reasons. However, in principle it should be possible to reconstruct multi-dimensional distributions as well. The variance optimization problem in that case would need to capture the covariance structure in gene expressions.

We begin by plotting the recovered vs original distributions for a couple more genes Nnat (Figure 4) and Cck (Figure 5). Since much more of the mass of these distributions is concentrated near 0, it makes it harder to recover these distributions when compared to Cpe. Essentially, Algorithm 1, and compressed sensing algorithms in general, work best when the original data is not too concentrated[8] and neither is it too dense.

We provide some details about the non-linear functions used in reconstructing the data. Suppose we can design probes that hybridize with two RNA strands as opposed to one, then the concentration of "doubly" hybridized product formed is $\sum_{i\in[N]} c_i^2$. Therefore, the mean and variance are proportional to $\sum_{i\in[N]} c_i^2$. So we can essentially measure quadratic functions of cellular concentrations using appropriate hybridization probes. We say $f(x) =$

[7] Note that the noise η can be reduced by appropriately averaging our observations before the sparse recovery, i.e., before step 6 in Algorithm 1. This averaging will however not reduce the reconstruction error at all, since it is compensated by an equal reduction in the minimum positive singular value in Equation 2.

[8] If the data is too concentrated or equivalently very sparse, a brute force search may be a better option.

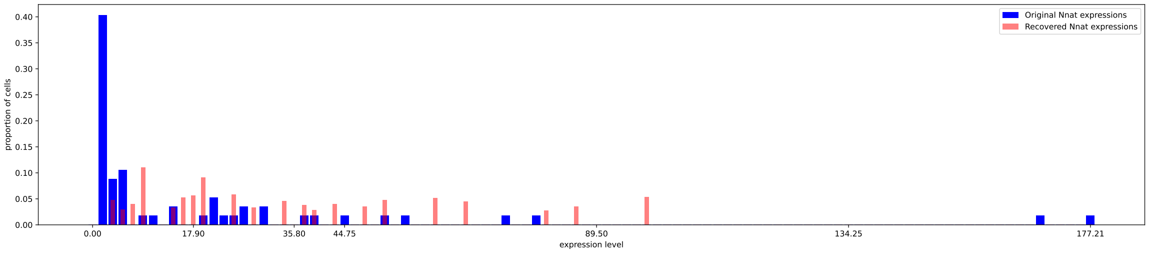


Figure 4: Recovered and original distributions using ℓ_1 minimization algorithm with a better optimized experiment design for Nnat $(N = 250, n = 100, m = 10, s = 20)$.

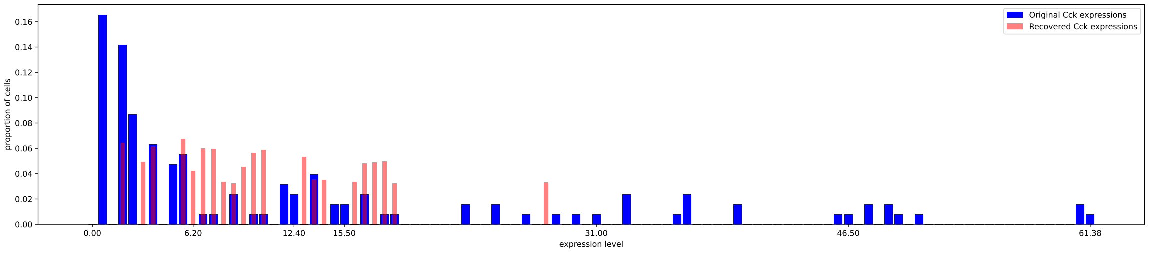


Figure 5: Recovered and original distributions using ℓ_1 minimization algorithm with a better optimized experiment design for Cck $(N = 250, n = 100, m = 10, s = 20)$.

x^2 is a *permissible function* and ensembles of the form:[9]

$$\begin{bmatrix} N(0,k'M_1^2) & N(0,k'(M_1+1)^2) & N(0,k'(M_1+2)^2) & \cdots \\ \vdots & \vdots & \ddots & \\ N(0,k'M_1^2) & N(0,k'(M_1+1)^2) & & N(0,k'M_2^2) \end{bmatrix}$$

lead to a *permissible measurement matrix* corresponding to the *permissible covariance matrix*: $\mathrm{Diag}(k'M_1^2, ..., k'M_2^2)$.[10]

Similarly we can measure third moments also. It is likely that we can't compute higher moments than three, as the chances of any meaningful amount of hybridization would be too small. On the other hand, we can also design probes that bind to our target RNA T or a background RNA B, but not both. If we know that the levels of B are constant in most cells then we can effectively measure using functions of the form $\frac{a}{b+c_T}$, where c_T is the concentration of our target RNA in the cell compartment, and a and b are constants depending on the background RNA used. They correspond to covariance matrices of the form $\mathrm{Diag}(\frac{a}{b+M_1}, \frac{a}{b+M_1+1}, ..., \frac{a}{b+M_2})$.

[9] We have assumed bins of width 1 for simplicity.

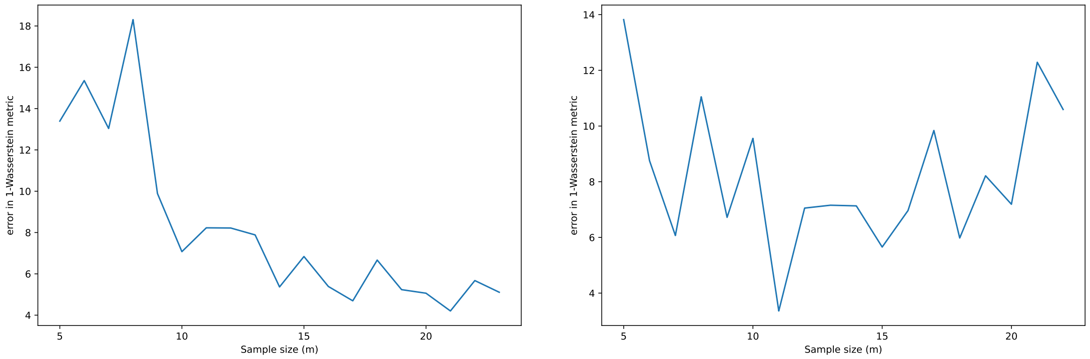
[10] Notation: Diag denotes the diagonal matrix with the specified diagonal.

Therefore, it is likely that a rich set of ”low degree” probes can be constructed. In our simulation, we used the following class of measurement functions:

$$F := \{x, x^2, \frac{a}{b+x}, \frac{a}{(b+x)(c+x)}, \frac{a'}{b'+x}, x^3\}.$$

Of course, nothing prevents us from repeating the same measurement twice, for example, F_1, F_1, F_1, F_1, F_1 which led to Figure 2. But, as noted, it will have diminishing returns as the smallest positive singular value (cf. Equation 2) will be smaller then. In particular, for our experiments with $m = 24$, we used the sequence of measurements: $F_1, F_2, ..., F_6, F_1, F_2, ..., F_6, F_1, F_2, ..., F_6, F_1, F_2, ..., F_6$.

In Figure 6, we plot the reconstruction error as a function of sample size (m). It is worth noting that 20 samples were enough for Cpe reconstruction to converge but not Cck reconstruction. Much more of the probability mass in the distribution of Cck is concentrated around 0 than for Cpe, which is probably what makes it harder to reconstruct.



(a) Error for Cpe

(b) Error for Cck

Figure 6: Error (measured by 1-Wasserstein distance between reconstructed and original distributions) as a function of sample size (m). Based on the trends, the error for Cpe behaves sub-linearly, so the upper-bound of $O(ms)$ on the error is likely not tight for Cpe.

Finally, just for the sake of comparison, we compare the ℓ_1 objective against an ℓ_2 objective. In other words, we replace the objective in Equation 1 by the ℓ_2^2 norm and make the same set of 24 measurements and reconstruct the distribution, keep the top 20 values and normalize (we do not average). The plot is shown in Figure 7. As is expected, ℓ_2 norm leads to a more even and spread out distribution. If we were to chose the 20 largest values in the reconstructed distribution, then we would be breaking many ”ties” arbitrarily and the resultant would still be skewed to higher values.

5. Discussion

This work proposes an algorithm to approximately reconstruct single cell distribution from a small number of aggregate measurements. On the algorithmic side, we propose an algorithm

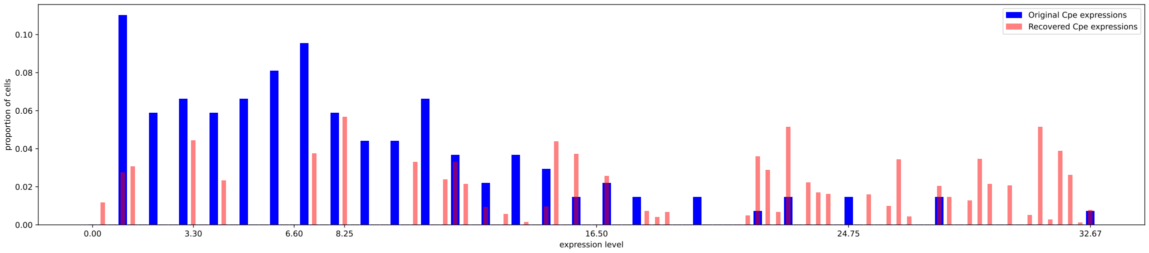


Figure 7: Plot of recovered and original distributions using ℓ_2 minimization.

and extend current compressed sensing results to show bounds for its performance and correctness. We use simulations using publicly available scRNA-seq data from Bastidas-Ponce et al. (2019) as the ground truth, to compare the original and approximately reconstructed distributions, where the latter is computed using our algorithm.

Limitations The major limitation of our work is our reliance on simulation using publicly available scRNA-seq values in place of our own experiments with in-situ hybridization. In order to verify that our proposed algorithm can actually reconstruct distributions with low error in practice, we need custom hybridization probes and other appropriate equipment. Although the overall method of making aggregate measurements is likely much simpler than carrying out current methods like FISH (for reasons explained in Subsection 2.2), the author currently does not have access to the resources to do the actual experiments.

References

Aimée Bastidas-Ponce, Leander Dony Sophie Tritschler, Katharina Scheibner, Marta Tarquis-Medina, Ciro Salinno, Silvia Schirge, Ingo Burtscher, Anika Böttcher, Fabian J Theis, Heiko Lickert, and Mostafa Bakht. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, 146(12): dev173849, 2019.

Volker Bergen, Marius Lange, Stefan Peidli, F Alexander Wolf, and Fabian Theis. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38:1408–1414, 2020.

Jameson Cahill, Xuemei Chen, and Rongron Wang. The gap between null space property and the restricted isometry property. *arxiv:1506.03040v2*, 2016.

Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008. URL https://www.sciencedirect.com/science/article/pii/S1631073X08000964.

Young Hwang Chang, Joe Gray, and Claire Tomlin. Exact reconstruction of gene regulatory networks using compressive sensing. *BMC Bioinformatics*, 15(400), 2014.

Xuemei Chen. Stability of compressed sensing for dictionaries and almost sure convergence rate for the kacmarz algorithm. *Diss. Vanderbilt University*, 2012.

Brian Cleary, Brooke Simonton, Jon Bezney, Evan Murray, Shahul Alam, Anubhav Sinha, Ehsan Habibi, Jamie Marshall, Eric S. Lander, Fei Chen, and Aviv Regev. Compressed sensing for highly efficient imaging transcriptomics. *Nature Biotechnology*, 39:936–942, 2021.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications.* Springer Verlag, New York, 1998. ISBN 0-387-98406-2.

Simon Foucart and Holger Rauhut. A mathematical introduction to compressive sensing. *Birkhäuser New York, NY*, 2296-5009:625, 2013.

Mohamed F. Haroon, Connor T. Skennerton, Jason A. Steen, Nancy Lachner, Philip Hugenholtz, and Gene W. Tyson. Chapter One - In-Solution Fluorescence In Situ Hybridization and Fluorescence-Activated Cell Sorting for Single Cell and Population Genome Recovery. In Edward F. DeLong, editor, *Microbial Metagenomics, Metatranscriptomics, and Metaproteomics*, volume 531 of *Methods in Enzymology*, pages 3–19. Academic Press, 2013. doi: https://doi.org/10.1016/B978-0-12-407863-5.00001-0. URL https://www.sciencedirect.com/science/article/pii/B9780124078635000010.

Linping Hu, Kun Ru, Li Zhang, Yuting Huang, Xiaofan Zhu, Hanzhi Liu, Anders Zetterberg, Tao Cheng, and Weimin Miao. Fluorescence in situ hybridization (FISH): an increasingly demanded tool for biomarker research and personalized medicine. *Biomarker Research*, 2(3), 2014.

Oren Jaspan, Roman Fleysher, and Michael L Lipton. Compressed sensing MRI: a review of the clinical literature. *British Journal of Radiology*, 88(1056), 2015.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(78):2241–2259, 2010. URL http://jmlr.org/papers/v11/raskutti10a.html.

6. Appendix

6.1. Proofs

6.1.1. Proof of Theorem 7

Proof Recall, from Equation 2, that

$$\|z_s(i) - x\|_1 \leq \frac{4\sqrt{2}\sqrt{N}}{(1-\gamma)\lambda(\Phi)} \|\Gamma x - \tilde{\Gamma} x\|_2, \tag{10}$$

where we have used: (1) $\sigma_s(x) = 0$ since x is s-sparse, and (2) $\eta = \|(V + \tilde{\Gamma})\Delta\|_2 = \|\Gamma x - \tilde{\Gamma} x\|_2$.

Therefore, squaring both sides of Equation 10 and averaging over the k terms implies:

$$\frac{1}{k}\sum_{i\in[k]} \|z_s(i) - x\|_1^2 \leq \left(\frac{4\sqrt{2}\sqrt{N}}{(1-\gamma)\lambda(\Phi)}\right)^2 \cdot \frac{1}{k}\sum_{i\in[k]} \|\Gamma(i)x - \tilde{\Gamma}(i)x\|_2^2. \tag{11}$$

By the Cauchy-Schwarz inequality and convexity of ℓ_1 norm, the LHS of Equation 11 is lower-bounded as:

$$\|x - \frac{\sum_{i\in[k]} z_s(i)}{k}\|_1^2 \leq \frac{1}{k} \sum_{i\in[k]} \|z_s(i) - x\|_1^2. \tag{12}$$

Since x is assumed s-sparse,

$$\|x - \tilde{p}\|_1 \leq \|x - \frac{\sum_{i\in[k]} z_s(i)}{k}\|_1. \tag{13}$$

The RHS of Equation 2 is a sum of k identical chi-square random variables and can be rewritten as:

$$\sum_{i\in[k]} \|\Gamma(i)x - \tilde{\Gamma}(i)x\|_2^2 = \mathrm{Var}(\Gamma x - \tilde{\Gamma}x). \tag{14}$$

The proof now follows by taking square roots, observing that $\|\Gamma v\|_2 \geq \frac{M_1}{M_2}\|v\|_2$ so that an appropriate RE condition is satisfied, and then finally using Lemma 6 to obtain the $O(ms)$ bound. ■

6.1.2. Proof of Theorem 8

Proof The proof modifies the first part of the proof of Theorem 1 in Raskutti et al. (2010). We begin with a recap of their theorem and corollary in our notation.

Theorem 9 (Theorem 1 in Raskutti et al. (2010)) *Let $\rho(\Upsilon) := \max_{i\in[n]} \Upsilon_{i,i}$. For any Gaussian random matrix $\Phi \in \mathbb{R}^{m\times n}$ with i.i.d. $N(0, \Upsilon)$ rows, there exist absolute constants c, c' such that:*

$$\frac{\|\Phi v\|_2}{\sqrt{m}} \geq \frac{1}{4}\|\Upsilon^{1/2}v\|_2 - 9\rho(\Upsilon)\sqrt{\frac{\log n}{m}}\|v\|_1, \tag{15}$$

with probability at least $1 - c'\exp(-cm)$.

As a corollary they essentially show:

Theorem 10 *Raskutti et al. (2010) Suppose that (s, α, γ)-RE condition with respect to Φ holds then, if*

$$m \geq c\frac{\rho^2(\Upsilon)(1+\gamma)^2}{\alpha^2} k \log n, \tag{16}$$

then Φ satisfies (s, γ)-NSP and thus Equation 2 holds for the ℓ_1-minimizer in Equation 1.

The proof of Theorem 10 from Theorem 9 remains unchanged in our case, so we focus on the changes required to the proof of Theorem 9. The proof of Theorem 9 in Raskutti et al. (2010) proceeds in three parts: (1) An upper-bound on $\mathbb{E}[M(r, \Phi)]$, where $M(r, \Phi) := 1 - \inf_{v\in V(r)} \frac{\|\Phi v\|_2}{\sqrt{m}}$, [11] (2) a sharp concentration result for $M(r, \Phi)$ around its expectation, and (3) a peeling argument to show the final high probability statement holds.

[11] Here $V(r) := \{v \in \mathbb{R}^n : \|\Upsilon^{1/2}v\|_2 = 1, \|v\|_1 \leq r\}$.

In order to derive the form of the objective function in Equation 9, we only need to non-trivially modify (1). The rest of the proof in Raskutti et al. (2010), steps (2) and (3) above, are concentration results which don't need any non-trivial modification.

The proof of (1) proceeds via Gordon's inequality, which is interesting in its own right. Our modification to the proof of (1), which closely resembles the original proof in Raskutti et al. (2010), follows.

For our purposes, $V(r)$ will need to be redefined with respect to a higher dimensional space. Let Υ denote the $mn \times mn$ block matrix, with diagonal blocks $\{\Upsilon_1, ..., \Upsilon_m\}$ and 0s elsewhere. Here each diagnoal block is a $n \times n$ covariance matrix in $\mathcal{M} = \{\Upsilon_1, ..., \Upsilon_m\}$. So Υ has the form:

$$\Upsilon = \begin{bmatrix} \Upsilon_1 & 0 & \dots \\ \vdots & \ddots & \\ 0 & & \Upsilon_m \end{bmatrix}$$

and redefine $V(r)$ as:

$$V(r) := \{v \in \mathbb{R}^n : \|\Upsilon^{1/2}(v, .., v)\|_2 = 1, \|v\|_1 \leq r, \},$$

where $(v, .., v)$ is simply a m times concatenation of v. Let $\rho(\{\Upsilon_1, ..., \Upsilon_m\}) := \max_j \sum_{i\in[m]} \Upsilon_i(i, j)$. We denote $\rho(\{\Upsilon_1, ..., \Upsilon_m\})$ by $\rho(\Upsilon)$ for brevity.

Recall that $M(r, \Phi)$:

$$M(r, \Phi) := 1 - \inf_{v\in V(r)} \frac{\|\Phi v\|_2}{\sqrt{m}} \tag{17}$$

Similar to Raskutti et al. (2010), we want to show:

$$\forall r > 0, V(r) \neq \emptyset \implies \mathbb{E}[M(r, \Phi)] \leq \frac{1}{4} + 3\rho(\Upsilon)\sqrt{\frac{\log n}{m}} r. \tag{18}$$

This is the counterpart of Lemma 1 in Raskutti et al. (2010) and constitutes step (1) above.

Gordon's inequality, in the form we need, states: For $(u, v) \in S^{m-1} \times V(r)$,[12] let $Y_{uv} := u^T \Phi v$ and $Z_{uv} := g^T u + h^T (\sum_{i\in[m]} \Upsilon_i^{1/2}) v$, for standard Normals $h \in N(0, I_{n\times n}), g \in N(0, I_{m\times m})$.[13] And if $\text{Var}(Y_{uv} - Y_{u'v'}) \leq \text{Var}(Z_{uv} - Z_{u'v'})$ for all $(u, v), (u', v')$ then:

$$\mathbb{E}[\sup_v \inf_u Y_{uv}] \leq \mathbb{E}[\sup_v \inf_u Z_{uv}]. \tag{19}$$

The usefulness of Gordon's inequality lies in the following observation:

$$\mathbb{E}[M(r, \Phi)] = 1 + \mathbb{E}[\sup_v \inf_u Y_{uv}] \leq 1 + \mathbb{E}[\sup_v \inf_u Z_{uv}].$$

Next, we need to verify that $\text{Var}(Y_{uv} - Y_{u'v'}) \leq \text{Var}(Z_{uv} - Z_{u'v'})$. This part follows exactly the proof in Raskutti et al. (2010) but with $\Sigma^{1/2}v$ substituted by $\left(\sum_{i\in[m]} \Upsilon_i^{1/2}\right) v$.

[12] S^{m-1} is the m-dimensional unit sphere.

[13] I denotes the identity matrix.

Finally, applying Gordon's inequality, similar to Raskutti et al. (2010), we get:

$$\begin{aligned}
\mathbb{E}[\sup_v \inf_u Y_{uv}] &\leq \mathbb{E}[\sup_v \inf_u Z_{uv}] &(20)\\
&= \mathbb{E}[\inf_u g^T u] + \mathbb{E}\sup_v h^T(\sum_{i\in[m]} \Upsilon_i^{1/2})v] &(21)\\
&= -\mathbb{E}[\|g\|_2] + \mathbb{E}[\sup_v h^T(\sum_{i\in[m]} \Upsilon_i^{1/2} v]. &(22)
\end{aligned}$$

Finally, we arrive at the crux of the difference between the proof of Lemma 1 in Raskutti et al. (2010) and our proof.

$$\sup_v |h^T(\sum_{i\in[m]} \Upsilon_i^{1/2})v| \leq \|v\|_1\|(\sum_{i\in[m]} \Upsilon_i^{1/2})h\|_\infty. \qquad (23)$$

The j^{th} co-ordinate of $(\sum_{i\in[m]} \Upsilon_i^{1/2})h$ is a mean zero Gaussian with variance exactly equal to our objective function in Equation 9: $\max_j \sum_{i\in[m]} \Upsilon_i(i,j)$. It is also the value of $\rho(\Upsilon)$ in the upper-bound on the sample size equation (Equation 16). The rest of our proof is very similar to the proof of Theorem 1 in Raskutti et al. (2010), and we refer the interested reader to their paper.

The conclusion is that for $m \simeq c\frac{\rho^2(\Upsilon)(1+\gamma)^2}{\alpha^2}k\log n$, a (s,α,γ)-RE condition holds with high probability. Note that an upper-bound on reconstruction error, that is directly proportional to m (similar to Theorem 7), holds and so minimizing $\rho(\Upsilon)$ is key to minimizing the reconstruction error. Thus the statement of Theorem 8. ■